

U.S. PATENT APPLICATION

Inventor(s): Christopher Andrew BARTON
Simon Neil REED
Martin Alan BROWN

Invention: PRE-APPROVAL OF COMPUTER FILES DURING A MALWARE
DETECTION

***NIXON & VANDERHYE P.C.
ATTORNEYS AT LAW
1100 NORTH GLEBE ROAD, 8TH FLOOR
ARLINGTON, VIRGINIA 22201-4714
(703) 816-4000
Facsimile (703) 816-4100***

SPECIFICATION

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

5

APPLICATION PAPERS

10

OF

15

CHRISTOPHER ANDREW BARTON

20

SIMON NEIL REED

25

AND

30

MARTIN ALAN BROWN

35

FOR

**PRE-APPROVAL OF COMPUTER FILES DURING A MALWARE
DETECTION**

BACKGROUND OF THE INVENTION

Field of the Invention

5 This invention relates to data processing systems. More particularly, this invention relates to malware detection within data processing systems, such as, for example, detecting computer viruses, computer worms, computer Trojans, banned computer files and the like.

Description of the Prior Art

10 The threat posed by malware, such as computer viruses, is well known and is growing. Computer viruses are becoming more common, more sophisticated and harder to detect and counteract. Computer systems and software for counteracting malware typically operate by seeking to identify characteristics of known malware
15 within a computer file being checked. A malware signature file typically contains data for identifying many thousands of different types of computer virus, Trojan, worm etc, as well as some characteristics generally indicative of malware and against which a computer file will need to be checked. With the rapid increase in the number, complexity and size of computer files present on a computer and requiring checking,
20 the amount of processing required and accordingly time needed to conduct malware detection is disadvantageously increasing. In the case of an on-access scan which is performed before access is allowed to a computer file, the delay introduced by first scanning that computer file for the presence of malware can introduce a noticeable and disadvantageous delay in the responsiveness of the computer system. In the case
25 of an on-demand scan where the entire contents of a computer are checked for malware, this check can take many minutes to perform and render the computer unusable for other purposes during this time.

30 One technique for speeding up malware detection that has previously been used is only to scan types of file which are executable. Potentially executable file types were previously restricted to relatively few types, such as EXE file types and COM file types. However, with the advent of more complex files and structures within files, it is now difficult to safely assume that a particular file type cannot contain any executable content and accordingly cannot contain malware.

Furthermore, as well as requiring a larger number of types of file to be subject to scanning, if not all file types, the increased complexity of the structures within files results in more processing being required to unpack and unravel those structures in order to effectively detect any malware which may be present within those computer files.

It is known from United States Patent US-A-6,021,510 to provide an anti-virus accelerator which when a file is examined for an initial time and found to be clean, then a hash value for each scanned sector for that file can be stored. Upon a subsequent attempt to scan that file, the file sectors which were examined in the initial scan can be examined again and their hash values recalculated and compared with the stored hash values. If the hash values match, then the sector can be considered to be unaltered and still clean.

The paper "A Cryptographic Checksum For Integrity Protection" published in Computers & Security, Volume 6, 1987, pages 505-510 by F. Cohen describes a cryptographic checksum technique for verifying the integrity of information in a computer system with no built in protection.

SUMMARY OF THE INVENTION

Viewed from one aspect the present invention provides a computer program product carrying a computer program operable to control a computer to detect malware within a computer file, said computer program comprising:

identifying code operable to identify said computer file as potentially being a specific known malware free computer file;

determining code operable to determine one or more attributes of said computer file; and

comparing code operable to compare said one or more attributes determined from said computer file with corresponding stored attributes of said specific known malware free computer file; wherein

if said attributes match, then confirming said computer file as being said specific known malware free computer file; and

if said attributes do not match then performing further malware detection processing upon said computer file.

5 The present invention recognises that there are some computer files which are highly likely to be present on many different computers and installations. As an example, the Windows operating system produced by Microsoft Corporation (TM) is widely used on an overwhelming majority of personal computers in the business environment. This operating system includes many large and complex files which are present on all such computers. Some of these computer files take a sufficiently
10 disadvantageous degree of processing to malware scan that it instead becomes worthwhile to specifically check and identify a computer file as being a particular common computer file that is known to be malware free and would otherwise consume significant resources to be the subject of malware detection. Surprisingly, effectively pre-approving a relatively small number of computer files once they have
15 been positively identified as being those computer files can make a significant impact upon the overall malware detection speed and more than compensate for the additional complexity within the malware scanner which is needed to check for pre-approval. This technique runs counter to the general prejudice in the malware detecting field where it is considered that the huge variety of different computer
20 programs which may be stored and used on a computer necessitates a generic approach to malware detection whereby all the computer files need to be checked for all of the relevant different types of malware with which they may be infected or to which they may correspond.

25 It will be appreciated that in identifying a computer file as potentially being one of the specific known malware free computer files a variety of different characteristics and/or parameters associated with that computer file may be utilised. Advantageously, these include one or more of the file name, storage location and file size of the computer file concerned. These characteristics tend to be strongly
30 indicative of a particular computer file being one of the candidates for pre-approval.

 Whilst the technique could be used to pre-approval only a single specific known malware free file, such as a file which was otherwise particularly time

consuming to process, the technique is particularly well suited when a plurality of different specific known malware free computer files are checked against in the pre-approval process.

5 Once a computer file has been identified as potentially being a specific known malware free computer file, the attributes that may be calculated for it or detected within it in order to confirm that it has not been altered in any way include calculating a checksum from a portion, portions or all of the computer file, such as a MD5 checksum, checking the content of a specific portion or portions against known
10 content at those locations and the like. These techniques are effective in ensuring that the candidate computer file has not been tampered with and yet are quick to perform.

 If a computer file is not identified as a pre-approved computer file, then normal malware detection processing may be proceeded performed in which one or
15 more characteristics corresponding to known malware files are detected, such as from a malware signature file.

 The present technique is particularly well suited for pre-approval of specific known malware free computer files being one of an operating system file, a help file
20 and a malware detection software file itself. Such computer files are highly likely or certain to be present within a computer utilising the present technique and yet can have a large size and a complex structure which would otherwise consume considerable resources when the subject of malware detection.

25 It will be appreciated that the malware being detected can take a wide variety of different forms, including a computer virus, a computer worm, a computer Trojan, a banned computer file and a computer file containing banned data.

 Viewed from another aspect the present invention provides a method of
30 detecting malware within a computer file, said method comprising the steps of:
 identifying said computer file as potentially being a specific known malware free computer file;

 determining one or more attributes of said computer file; and

comparing said one or more attributes determined from said computer file with corresponding stored attributes of said specific known malware free computer file; wherein

if said attributes match, then confirming said computer file as being said
5 specific known malware free computer file; and

if said attributes do not match then performing further malware detection processing upon said computer file.

Viewed from a further aspect the present invention provides apparatus for
10 detecting malware within a computer file, said apparatus comprising:

identifying logic operable to identify said computer file as potentially being a specific known malware free computer file;

determining logic operable to determine one or more attributes of said
computer file; and

15 comparing logic operable to compare said one or more attributes determined from said computer file with corresponding stored attributes of said specific known malware free computer file; wherein

if said attributes match, then confirming said computer file as being said
specific known malware free computer file; and

20 if said attributes do not match then performing further malware detection processing upon said computer file.

The above, and other objects, features and advantages of this invention will be apparent from the following detailed description of illustrative embodiments which is to
25 be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a computer storing a plurality of different computer files;

30 Figure 2 illustrates the directory structure of a computer and the location and characteristics of certain computer files which may be subject to pre-approval;

Figure 3 illustrates pre-approval data which is used to identify a file as potentially being a specific known malware free computer file and then to confirm the identity of that computer file;

5 Figure 4 is a flow diagram illustrating pre-approval processing; and

Figure 5 is a diagram illustrating the architecture of a general purpose computer which may be used to implement the above techniques.

10 **DESCRIPTION OF THE PREFERRED EMBODIMENTS**

Figure 1 illustrates a computer 2 including a hard disk drive 4. The computer 2 includes malware detection software (AV software) which can perform both on-access and on-demand scanning. Also illustrated as stored upon the hard disk drive 4 are an operating system and an application program such as Microsoft Word word processing,
15 which is commonly found upon many computers 2.

Figure 2 schematically illustrates the directory structure used by the hard disk drive 4 to store the computer files. As illustrated, various directories and subdirectories are provided. Particular files which may be subject to pre-approval have known file
20 names and sizes. They are also typically stored within the same or a determinable relative location within the directory structure from system to system. For example, certain computer files might always be stored in the root directory, certain computer files always stored in a WINNT directory, an OFFICE directory or the like. The location may also be determined by an existing configuration setting, such as a registry entry,
25 configuration file, etc. The combination of file name, file type, file size and file location may be used to identify a particular computer file as potentially being a specific known malware free computer file which has been pre-approved.

Figure 3 schematically illustrates a table of data corresponding to several
30 different pre-approved computer files. For each pre-approved computer file, data is stored indicating file name, size and location, compared with that of a candidate computer file to identify that candidate computer file as potentially corresponding to a pre-approved computer file. Also stored is an MD5 checksum for the specific known

malware free computer file is calculated from a portion, portions, or all of the specific known malware free computer files. The same checksum is calculated for the candidate computer file and if the stored checksum and the calculated checksum match, then the candidate computer file is positively identified as being the pre-approved computer file and may be passed as clean without further malware scanning. It will be appreciated that as an alternative to checksum calculation, the direct content of the computer file at certain portions or a portion may be compared and checked in order to positively identify the candidate computer file.

In Figure 3, three pre-approved computer files are illustrated. In practice it has been found that using the present pre-approval technique in relation to what are otherwise the worst offending top 100 computer files in terms of the slowness of their malware detection scanning and the frequency with which they occur can overall yield significant improvements in malware detection processing speed, such as up to a 30% improvement. This is dramatic and highly significant. The extra time required for the pre-approval processing check and the additional storage requirement within the virus signature file is more than compensated for by the performance improvement achieved.

Figure 4 is a flow diagram schematically illustrating the pre-approval processing check. At step 6 the system waits until a computer file to scan is received into a scanning queue. At step 8, the file name, file location and file size of the candidate computer file are read. At step 10 these read parameters are compared with a list of pre-approved files stored within the computer virus definition data. Step 12 determines whether a match occurred. If there was no match, then processing proceeds to step 14 at which standard malware detection processing is conducted following which the AV scan result is output at step 15. If a match did occur at step 12, then the computer file is a candidate to be treated as a specific known malware free computer file which can be approved without detailed malware detection. However, before the computer file is actually so approved, it must be subject to a check to determine it has not been tampered with.

At step 16 an MD5 checksum for the candidate computer file is calculated. At step 18 this calculated checksum is compared with the stored checksum determined for

the specific known malware free computer file against which a match was detected at step 12. At step 20 if the checksums match, then processing proceeds to step 22 whereby the candidate computer file can be indicated as being clean without detailed malware detection needing to be performed. If the checksums did not match, then
5 processing proceeds from step 20 to step 14 where full malware detection is performed.

It will be appreciated that the processing steps illustrated in Figure 4, other than that of step 14, are relatively rapid compared to the long amount of time needed to conduct the standard malware detection scanning in step 14. Thus, a candidate computer
10 file will either be identified as an unaltered pre-approved computer file thereby terminating further processing requirements for that computer file or relatively rapidly handed on to the standard malware scanning system for malware detection.

Figure 5 schematically illustrates a general purpose computer 200 of the type that
15 may be used to implement the above described techniques. The general purpose computer 200 includes a central processing unit 202, a random access memory 204, a read only memory 206, a network interface card 208, a hard disk drive 210, a display driver 212 and monitor 214 and a user input/output circuit 216 with a keyboard 218 and mouse 220 all connected via a common bus 222. In operation the central processing unit
20 202 will execute computer program instructions that may be stored in one or more of the random access memory 204, the read only memory 206 and the hard disk drive 210 or dynamically downloaded via the network interface card 208. The results of the processing performed may be displayed to a user via the display driver 212 and the monitor 214. User inputs for controlling the operation of the general purpose computer
25 200 may be received via the user input output circuit 216 from the keyboard 218 or the mouse 220. It will be appreciated that the computer program could be written in a variety of different computer languages. The computer program may be stored and distributed on a recording medium or dynamically downloaded to the general purpose computer 200. When operating under control of an appropriate computer program, the
30 general purpose computer 200 can perform the above described techniques and can be considered to form an apparatus for performing the above described technique. The architecture of the general purpose computer 200 could vary considerably and Figure 5 is only one example.

Although illustrative embodiments of the invention have been described in detail herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to those precise embodiments, and that various changes and modifications can be effected therein by one skilled in the art without departing from the scope and spirit of the invention as defined by the appended claims.

5